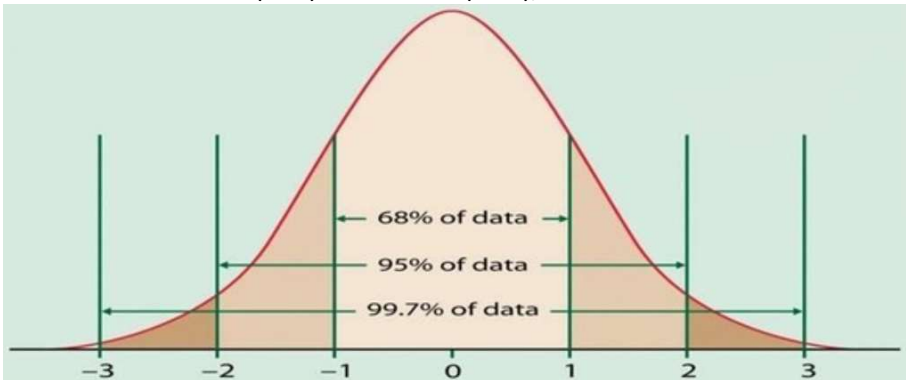


Statistics Basics

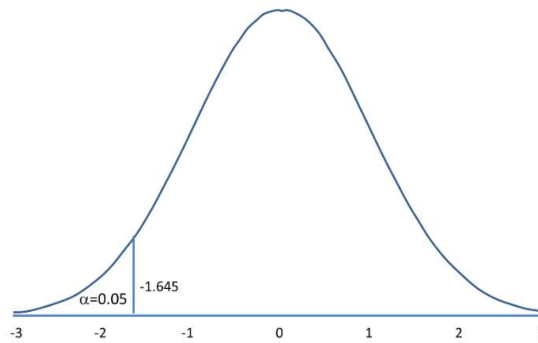
Saturday, July 8, 2017

9:50 AM

Topic	Description
Data	<ul style="list-style-type: none"> Collected observations. Data can be continuous(ex: product price) or categorical (ex: gender) Data helps us to understand the relationships if any between events Data helps to predict future behavior Visualizing the data helps to see the patterns that can't be seen in a table
Measuring Data	<ul style="list-style-type: none"> Nominal - Predetermined Categories. This can't be sorted. Ex: Gender Ordinal - Can be sorted. Lacks scale. Ex: Survey responses - often, sometimes, never etc Interval - Provides scale but lacks a 'zero point'. Ex: Celsius temperature Ratio - Values have a true zero point. Ex: Age, weight, salary etc
Measurements of central tendency	<ul style="list-style-type: none"> These measures provide the location of data Fail to describe the shape of the data <ul style="list-style-type: none"> Mean or average = sum of all the values/number of values <ul style="list-style-type: none"> Outliers will influence in calculating mean value Median is middle value <ul style="list-style-type: none"> Median is much closer to the most values in the series Mode is most common occurring value
Measurements of dispersion	<ul style="list-style-type: none"> Range - It is difference between maximum and minimum value Variance - sum of square distances from the each point to the mean. This differs between sample and population variance subject to Bessels correction (n-1). This helps to understand how dispersed are these values. Resulting value is in squared value, which does not provide correct inference. So that why we using standard deviation <p>SAMPLE VARIANCE: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$</p> <p>POPULATION VARIANCE: $\sigma^2 = \frac{\sum(X-\mu)^2}{N}$</p> Standard Deviation - it is square root of variance. Since units are same as sample it is provide correct inference. <div style="border: 1px solid black; padding: 10px; width: fit-content; margin: 10px auto;"> $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$ </div>
Measurements Quartiles	<ul style="list-style-type: none"> Every value in the data is used, no aggregation methods like mean or variance Divides series into 4 parts and will have 3 quartiles Then box plot is plotted to check the distribution Inter Quartile range (IQR) is used to find the outlier To find the outliers a fence will be created at 1.5 IQR. Anything after that range is outlier
Covariance, Correlation	<ul style="list-style-type: none"> Correlation is finding relation between two variables Scatter plots best to uncover correlation Covariance is comparing variances of two variables First step is to match scale of two variables $cov(X, Y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$ Pearson correlation coefficient In order to normalize values from two distribution we use $cov(X,y)/[(std\ dev\ X)(std\ dev\ y)]$ Correlation value falls between +1 and -1, +1 is total -ve linear correlation, 0 is no linear

	<p>correlation, -1 is total negative linear correlation</p> <ul style="list-style-type: none"> Correlation does not necessarily mean the causality. We need to look into other variables and need to do other statistical analysis for causality
Probability	<ul style="list-style-type: none"> Probability is a value between 0 and 1 that a certain event will occur Each trial is an experiment and each mutually exclusive outcome is an event
Permutations	<ul style="list-style-type: none"> Permutation is arrangement of the objects in all possible ways Total permutations of set n is n! Permutations taken r at a time given set n (no repetition) is $n!/(n-r)!$ Permutations taken r at a time given set n (repetition) is n^r
Combinations	<ul style="list-style-type: none"> Combinations are unordered arrangements of objects Number of combinations of a set of n objects taken r at a time (no repetition) $n!/r!(n-r)!$ Number of combinations of a set of n objects taken r at a time (repetition) $(n+r-1)!/r!(n-1)!$
Bayes Theorem	<ul style="list-style-type: none"> Bayes Theorem is used to determine the probability of a <i>parameter</i>, given a certain event. The general formula is: $P(A B) = \frac{P(B A) \cdot P(A)}{P(B)}$
Distributions	<p>Distributions describes all of the probable outcomes of a variable</p> <p>In discrete distribution, sum of all the individual probabilities must be equal to 1</p> <p>In continuous distribution, the area under the probability curve equal to 1</p>
Discrete probability distributions OR Probability Mass Functions	<p>Uniform distribution - Each of the value will have same probability and sum is equal to 1</p> <p>Binomial Distribution - It is number of success per number of trials. If it trial is 1, then there will 2 possible outcomes and sum is equal to 1. This can be done for n trials</p> <p>Poisson Distribution- is number of success per any continuous unit</p>
Continuous probability distributions OR Probability density functions	<p>Exponential Distribution.</p> <p>Beta Distribution</p> <p>Normal Distribution (Gaussian or bell curve)</p> <ul style="list-style-type: none"> In this area under curve is equal to 1. This is always symmetrical Probability of specific outcome is zero We can find probabilities over specified interval or range of outcomes Mean, median and mode are always equal Lower tail, upper tail For standard normal (or Z) distribution(SND), mean = 0 and std dev =1  <ul style="list-style-type: none"> We can take any Normal distribution and standardize it using Z-score to make SND Z-score is $(x - \text{mean}) / \text{std dev}$

	<ul style="list-style-type: none"> • It indicates how many standard deviations an element is from the mean • It enables us to compare two scores that are from different normal distributions. The Z-score does this by standardizing scores in a normal distribution to z-scores in standard normal distribution. • Basically we will be using z-score to calculate percentile by looking at Z-table • Using python: <pre>from scipy import stats p = 0.95 Stats.norm.ppf(p) 1.644853 # this is Z-score</pre>
Statistics	<ul style="list-style-type: none"> • It is science of data. Using statistics, we are trying to infer whole population from those in a sample • Population is every member of a group • Sample is a subset of the population • Usually >30% from the population is good sample
Sampling	<ul style="list-style-type: none"> • Random- sample with equal chance • Stratified Random - sample will have adequate group representation • Clustering - sample selected from particular clusters • Central Limit Theorem - It states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population and it will be normally distributed
Hypothesis Testing	<ul style="list-style-type: none"> • It is a statistical method that is used in making statistical decisions using sample data • There are two ways to do Hypothesis Testing • Traditional Method <ol style="list-style-type: none"> 1. From the problem statement <ul style="list-style-type: none"> ▪ We will know mean, std. dev of overall population ▪ Also we will have sample details of number of runs and mean ▪ If significance level is not given, we will assume it as 0.05 2. We assume the null hypothesis (The default position of null hypothesis states that there is no statistics significance between 2 variables) and alternative hypothesis. Null and alternate are mutually exclusive 3. Determine the test type - left, right or two tail depending on null hypothesis we assumed. Level of significance is the area inside the tails of our null hypothesis <div data-bbox="613 1354 1166 1478" data-label="Figure"> <p>left tail right tail two tail</p> </div> <ul style="list-style-type: none"> ▪ If $\alpha = 0.05$ and alternative hypothesis is less than the null, left-tail of probability curve has an area of 0.05

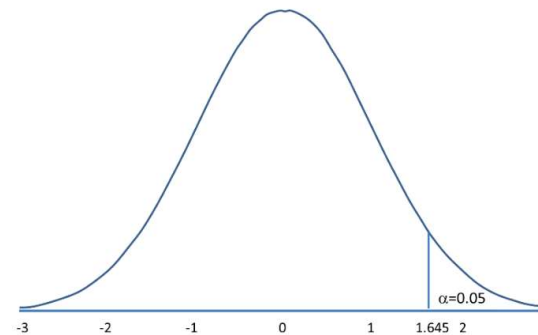


Rejection Region for Lower-Tailed Z Test ($H_1: \mu < \mu_0$) with $\alpha = 0.05$

The decision rule is: Reject H_0 if $Z \leq -1.645$.

Lower-Tailed Test	
α	Z
0.10	-1.282
0.05	-1.645
0.025	-1.960
0.010	-2.326
0.005	-2.576
0.001	-3.090
0.0001	-3.719

- If $\alpha = 0.05$ and alternative hypothesis is more than the null, right-tail of probability curve has an area of 0.05

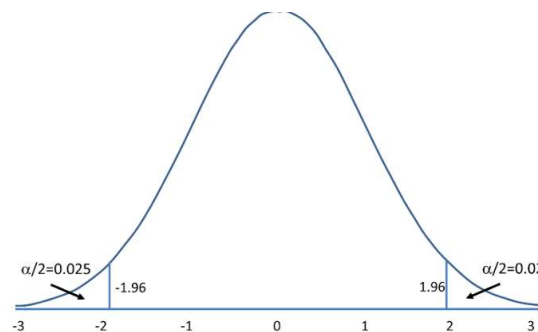


Rejection Region for Upper-Tailed Z Test ($H_1: \mu > \mu_0$) with $\alpha = 0.05$

The decision rule is: Reject H_0 if $Z \geq 1.645$.

Upper-Tailed Test	
α	Z
0.10	1.282
0.05	1.645
0.025	1.960
0.010	2.326
0.005	2.576
0.001	3.090
0.0001	3.719

- If $\alpha = 0.05$ and alternative hypothesis is not equal to null, both-tails of probability curve share an area of 0.05



Rejection Region for Two-Tailed Z Test ($H_1: \mu \neq \mu_0$) with $\alpha = 0.05$

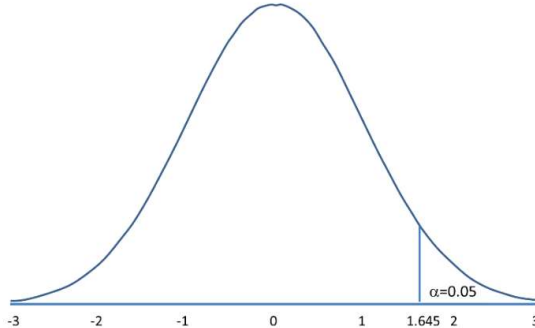
The decision rule is: Reject H_0 if $Z \leq -1.960$ or if $Z \geq 1.960$.

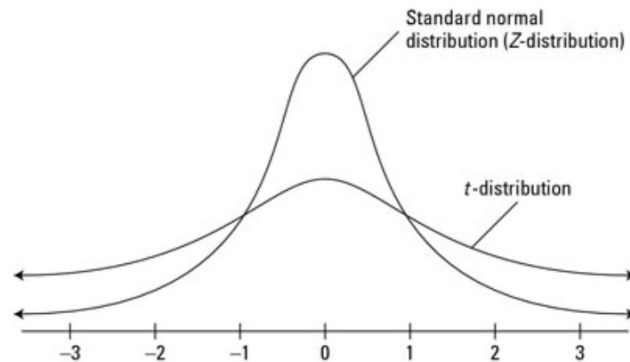
Two-Tailed Test	
α	Z
0.20	1.282
0.10	1.645
0.05	1.960
0.010	2.576
0.001	3.291
0.0001	3.819

- Now, We either calculate mean or proportion test to calculate test statistic
- Then we determine the Z-score based on level of significance in Z-table. These areas establish our critical values or Z-scores (if are using both tails)
- Compare the test statistic to the z-score
- Based on this either we reject (if null hypothesis not true) or fail to reject (if null hypothesis is true) null hypothesis
- If the null hypothesis is rejected, then we say that data support alternate hypothesis

- Academic Method(using P-value Test)

- We will follow above steps till step4, to calculate the test statistic
- Then we use this test statistic and will look up in z-table to find p-value
- Compare the p-value to the level of significance (α)

	4. Then will follow steps 7, 8 from above																				
Hypothesis test to find whether performance of the app increased or not	<div>1. Problem statement - Company got an app, it's current mean loading time is 'mew', its standard deviation is 'sigma'. They wanted to improve the performance of the loading time, so they asked development team to work on it. They finetuned it, after that they tested the app 'n' number of times and found that mean of the loading time is 'x-bar'. Development team is 95% confidence that performance is increased. Now company wanted to know whether the app's performance is increased or not.</div> <div>2. From the problem statement<ul style="list-style-type: none">▪ We know mean(mew), std. dev(sigma) of current app performance▪ Also we will have sample details of number of runs 'n' and mean 'x-bar'▪ Also development team has 95% confidence. That means significance level is 0.05 (If significance level is not given, we will assume it as 0.05)</div> <div>3. We assume null hypothesis is $h_0(x) \leq 0.05$ and alternative hypothesis $h_1(x) > 0.05$</div> <div>4. Determine the test type - left, right or two tail depending on null hypothesis we assumed. Since alternative hypothesis $h_1(x) >$ null hypothesis $h_0(x)$, we consider right tail<ul style="list-style-type: none">▪ If $\alpha = 0.05$ and alternative hypothesis is more than the null, right-tail of probability curve has an area of 0.05</div> <div><div><p>Rejection Region for Upper-Tailed Z Test ($H_1: \mu > \mu_0$) with $\alpha=0.05$</p><p>The decision rule is: Reject H_0 if $Z \geq 1.645$.</p></div><div><table><tr><th colspan="2">Upper-Tailed Test</th></tr><tr><th>α</th><th>Z</th></tr><tr><td>0.10</td><td>1.282</td></tr><tr><td>0.05</td><td>1.645</td></tr><tr><td>0.025</td><td>1.960</td></tr><tr><td>0.010</td><td>2.326</td></tr><tr><td>0.005</td><td>2.576</td></tr><tr><td>0.001</td><td>3.090</td></tr><tr><td>0.0001</td><td>3.719</td></tr></table></div></div> <div>5. Now, We calculate mean test to calculate 'test statistic' because we got mean values in our problem statement.</div> <div>6. Then we determine the Z-score based on level of significance in Z-table.<ul style="list-style-type: none">▪ Based on below table we know if significance level is 0.05 Z-score is 1.645</div> <div>7. Compare the 'test statistic' to the Z-score (1.645) calculated from significance level</div> <div>8. Based on this either we reject (if null hypothesis not true) or fail to reject (if null hypothesis is true) null hypothesis</div> <div>9. If the null hypothesis is rejected, then we say that data support alternate hypothesis</div>	Upper-Tailed Test		α	Z	0.10	1.282	0.05	1.645	0.025	1.960	0.010	2.326	0.005	2.576	0.001	3.090	0.0001	3.719		
Upper-Tailed Test																					
α	Z																				
0.10	1.282																				
0.05	1.645																				
0.025	1.960																				
0.010	2.326																				
0.005	2.576																				
0.001	3.090																				
0.0001	3.719																				
Type I and Type II errors	<div>Type I - Rejected the null hypothesis but should have supported it</div> <div>Type II - Fail to reject the null hypothesis but should have rejected it</div> <div>Other way to put this is</div> <div>Type I is predicting YES where actual is NO (False Positive)</div> <div>Type II is predicting NO where actual is YES (False Negative)</div> <div><table><tr><td></td><td>Predicted: NO</td><td>Predicted: YES</td><td></td></tr><tr><td>n=165</td><td></td><td></td><td></td></tr><tr><td>Actual: NO</td><td>TN = 50</td><td>FP = 10</td><td>60</td></tr><tr><td>Actual: YES</td><td>FN = 5</td><td>TP = 100</td><td>105</td></tr><tr><td></td><td>55</td><td>110</td><td></td></tr></table></div> <div>Basic Terminology:</div> <div><ul style="list-style-type: none">• True Positives (TP)• True Negatives (TN)• False Positives (FP)• False Negatives (FN)</div>		Predicted: NO	Predicted: YES		n=165				Actual: NO	TN = 50	FP = 10	60	Actual: YES	FN = 5	TP = 100	105		55	110	
	Predicted: NO	Predicted: YES																			
n=165																					
Actual: NO	TN = 50	FP = 10	60																		
Actual: YES	FN = 5	TP = 100	105																		
	55	110																			
T-Distribution	<ul style="list-style-type: none">• T-Distribution curve is almost similar to standard normal (Z) distribution• t-distribution is shorter at the center and got slightly fatter tails compared to Z or Standard normal distribution. Comparing Z-Distribution with T-Distribution																				



- When the degrees of freedom increases, t-distribution becomes more like z-distribution
- t-distribution is used typically to study the mean of the population rather than on individuals within a population. Also it is used when we have small data sample and when we don't know standard deviation (sigma)
- The z-distribution applies when the variance is fixed. Student's t-distribution applies when the variance varies
- The t-distribution with one degree of freedom is also known as the Cauchy distribution
- Types - one-sample, two-sample,
- Steps to perform the two sample T-Distribution
 1. First get the two sample and compare their means
 2. Now we assume the null hypothesis and alternate hypothesis based on about means
 3. Calculate the degrees of freedom
 4. Calculate the variance of two samples
 5. Calculate t-value base on above variances, means, degrees of freedom
 6. Based on degrees of freedom, significance level we will find critical value from t-table
 7. Then we compare above calculated t-value with t-value we got from t-table
 8. Based on this either we reject (if null hypothesis not true) or fail to reject (if null hypothesis is true) null hypothesis
 9. If the null hypothesis is rejected, then we say that data support alternate hypothesis
- Steps using python

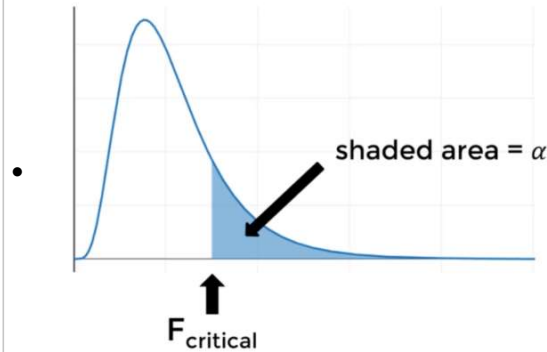

```

      From scipy.stats import ttest_ind
      A = [...]
      B = [...]
      Statistics_value = ttest_ind(a,b).statistic
      t_value = ttest_ind(a,b).pvalue/2 # this is for two-sample test
      Compare statistics_value with t_value
      
```

Analysis of Variance and F-distribution

- It is a collection of statistical models and their associated variance used to analyze the differences among groups means in a sample
- ANOVA is useful for comparing (testing) three or more group means for statistical significance
- We have one way ANOVA and two way ANOVA.
- In one way ANOVA, we will have one independent variable
- In two way ANOVA, we will have two independent variable at same time
- Steps to calculate one way ANOVA
 - Assume null hypothesis
 - Calculate the sample means
 - Calculate over all means
 - Calculate variance between groups
 - Calculate variance within groups
 - Calculate F-value = Variance between groups/Variance within groups

- Find f critical value for alpha (level of significance) = 0.05
- Compare Calculated F-value with Critical value to see if we need to reject or fail to reject null hypothesis
- F-distribution will be something like this ..



- In python
`from scipy import stats`
`stats.f.ppf(1-.05, dfn=2, dfd=27)`

Difference between Z, T, F, chi square test

Z-Test	T-Test	F-Test
Used for testing the mean of a population versus a standard, or comparing the means of two populations	Used for testing the mean of one population against a standard or comparing the means of two populations	Used to compare 2 populations variances
Sample size (n>30)	If you do not know the populations standard deviation or when you have a limited sample (n < 30)	The samples can be any size
Samples <3	Samples <3	Sample 3 or more

Chi-Square Analysis

It is used to compare the observed frequencies in a table to the expected frequencies

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

In Python

From `scipy.stats` import `chi2`

`Chi2.isf(alpha, degrees of freedom)`